Matlab code for Problem 4

```
X_train = csvread('X_train.csv');
y_train = csvread('label_train.csv');
y_test = csvread('label_test.csv');
X_test = csvread('X_test.csv');

setNum = size(X_test,1);
setSize = size(X_train,2);
N1 = length(find(y_train));
N0 = length(find(~y_train));
p_pre1 = zeros(setNum,1);

sumX1 = sum(X_train.*repmat(y_train,1,setSize),1);
sumX0 = sum(X_train.*repmat(1-y_train,1,setSize),1);
log_cX = sum((sumX0+1)*(log(N0+1)-log(N0+2))) -
sum((sumX1+1)*(log(N1+1)-log(N1+2)));
log_cN = log(N1+2)-log(N0+2);

for k = 1:setNum
    log_factor1 = 0;
    log_factor0 = 0;
    for i = 1: 54
        if X_test(k,i) ~= 0
            log_factor1 = log_factor1 +
sum(log(sumX1(i)+1:sumX1(i)+X_test(k,i)));
            log_factor0 = log_factor0 +
sum(log(sumX0(i)+1:sumX0(i)+X_test(k,i)));
        end
    end
    log_fx = (sum(X_test(k,:))-1)*log_cN + log_factor0 - log_factor1;
    p0_div_p1 = exp(log_fx + log_cX + log(N0+1)-log(N1+1));
    p_pre1(k) = 1/(1+p0_div_p1);
end
y = (p_pre1 > 0.5);

r = (y == y_test);
s_s = length(find(y.*r));
s_n = length(find(y.*~r));
n_n = length(find(~y.*r));
n_s = length(find(~y.*~r));
cNames = {'classified spam', 'classified non_spam'};
rNames = {'spam', 'non-spam'};
data = [s_s n_s;n_s n_n];
classified_spam = [s_s;s_n];
classified_non_spam = [n_s;n_n];
table(classified_spam,classified_non_spam,'RowName',rNames)

m = find(~r);
figure;
for i = 1:3
    p_pre1(m((i)))
    plot(X_train(m(i),:),'-*');
    hold on;
```

```
end
plot((sumX1+1)/(N1+1))
hold on;
plot((sumX0+1)/(N0+1))
legend('sample1','sample2','sample3','E1','E0');
title('misclassified')

[~,I] = sort(abs(p_pre1-0.5));
figure;
 for i = 1:3
    p_pre1(I((i)))
    plot(X_train(I(i),:),'-*');
    hold on;
end
plot((sumX1+1)/(N1+1))
hold on;
plot((sumX0+1)/(N0+1))
legend('sample1','sample2','sample3','E1','E0');
title('three most ambiguous')
```